Journal of Ethics and Emerging Technologies

IEET

Book Review

One, none, one hundred thousand AIs

Pelillo, M., & Scantamburlo, T. (Eds.). (2021). *Machines We Trust: Perspectives on Dependable AI*. MIT Press.

Vincenzo Politi [1]

1. CEHIC - Center for the History of Science, Universitat Autónoma de Barcelona, 08193, Bellaterra (Barcelona, Spain). vin.politi@gmail.com

Abstract: Like many innovative technologies, AI possesses a transformational power: its implementation in society is not a neutral additive process, but it may alter in significant ways various social and cultural dynamics. Socio-ethical concerns led to the demand of designing AI devices that can be 'trusted'. The recent publication of *Machines we trust* provides novel opportunities to discuss some socio-ethical issues arising from human-AI interactions. After defining the concepts of trust, trustworthiness, and reliability, and explaining in which sense it is possible to talk about 'trustworthy AI', I focus on two chapters of the volume that consider some concrete applications of AI. I conclude by suggesting that, instead of considering the different contributions to the volume in isolation with respect to one another, it may be illuminating to compare and contrast them. Such a way of reading the book leads us to question whether it is still possible to talk about trustworthy AI 'in general' or whether the discussion about the socio-ethical issues posed by AI should proceed in a piece-meal case-by-case fashion.

Keywords: Artificial Intelligence; Trust; Social Responsibility; Unintended consequences; Experiments.

Citation: Politi, Vincenzo. 2022. "One, no one, one hundred thousand AIs". *Journal of Ethics and Emerging Technologies* 32: 2. https://doi.org/10.55613/jeet.v32i2.120

Received: 07/11/2022
Accepted: 15/11/2022
Published: 31/12/2022

Publisher's Note: IEET stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Like many innovative technologies, artificial intelligence (AI) possesses a transformational power: its implementation in society is not a neutral additive process, but it may alter in significant ways various social and cultural dynamics. Socio-ethical concerns led to the demand of designing AI devices that can be 'trusted'. For example, a high-level expert group set up by the European Commission has recently published an official guideline for 'trustworthy AI' (European Commission 2019). The document considers the multiplicity of actors involved in the production, implementation, and use of artificial intelligence (i.e., 'developers', 'deployers' and 'end-users'). It also provides a non-exhaustive list of requirements for trustworthy AI, such as:

human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and fairness, societal and environmental wellbeing, accountability. The document is one of the latest contributions to the governance and policy literature on AI. While it is evident that AI must be designed and implemented responsibly, however, it may not be entirely clear whether it is appropriate to talk about its 'trustworthiness'. The recent publication of *Machines we trust: perspectives on dependable AI* (Pelillo and Scantamburlo 2021; from now on: *MwT*) provides novel opportunities to discuss the socio-ethical issues arising from human-AI interactions as well as the sense of talking about 'trustworthy AI'.

In the next section, I briefly define the concepts of trust, trustworthiness, and reliability, and I also summarize the main contents of *MwT*. In the third and fourth section, I focus on two chapters: one about AI in healthcare and one about autonomous robotics. I conclude by suggesting that, instead of considering the different contributions to *MwT* in isolation with respect to one another, it may be illuminating to compare and contrast them. Such a way of reading the book leads us to question whether it is still possible to talk about trustworthy AI 'in general' or whether the discussion about the socio-ethical issues posed by AI should proceed in a piece-meal case-by-case fashion.

2. Trust, reliability, and social relations

Roughly speaking, we trust people if we have good reasons to think that they will tell us the truth, that they will act consistently to what they have expressed and do so in a non-hurtful way. People we trust are 'trustworthy' and we usually tend to treat them differently from those we do not trust. Trustworthiness, however, is not infallibility: a trustworthy person may make mistakes. This is why trust is not blind faith: we know that there exists the risk that trustworthy people may tell us something false or may behave in a hurtful way. Whether someone made a genuine and unintended mistake or, instead, betrayed our trust can be assessed by asking for an explanation. Together, trust and trustworthiness are at the basis of social relations.

Although they are often used interchangeably in our daily language, philosophers conceptually distinguish 'trustworthiness' and 'reliability' (Baier 1986). The latter concept describes a property possessed by inanimate objects, such as tools and devices. Something is reliable when we have good reasons to think that it will perform consistently well in carrying out a specific task (i.e., we rely on the hammer to drive the nail on the wall). The concept of reliability does not possess the same complex mixture of epistemological and ethical elements characterizing trustworthiness. An instrument is reliable when it does its work. Unreliable tools will be just deemed as faulty: they do not make mistakes despite their good intentions, nor are they trying to deceive us. Moreover, we cannot ask them the reasons behind their misbehavior: in fact, tools do not have 'reasons'. In short, while people whom we trust may potentially betray us, inanimate tools may only disappoint us (Holton 1994, Wright 2010, McLeod 2015).

In the light of these conceptual distinctions, one may wonder to what extent it is legitimate to talk about 'trustworthy AI', rather than just 'reliable AI'. On the one hand, based on the current state of research and on what we know, AI does not possess a consciousness and it is indeed implemented in inanimate tools and devices. On the other hand, however, it would not be

entirely correct to speak of reliability in the case of AI. What distinguishes AI from other technologies, in fact, is the kind of *relation* we have with it. While we use tools to perform specific tasks, AI plays increasing roles in our own reasoning and decision-making processes. As such, AI is not designed to just 'perform *x*' but, rather, to 'perform *x* in a way that is beneficial, non-harmful, just, transparent, and (at least to some extent) explainable'. For this reason, even though they are inanimate unexplainable black boxes, AI devices cannot be treated just as simply tools we can rely on and should be made the most trustworthy as possible.

MwT delves into this 'hybrid' character of AI and it may not be a coincidence that its subtitle mentions the concept of 'dependability', that somehow comprises both trustworthiness and reliability. The volume collects the contributions of experts from different fields ranging from the electronic and engineering sciences to the humanities. Commendably, the editors have managed to assemble an expert group that is not only *interdisciplinary*, but also truly *international*, with contributors coming from, and being based in, Belgium, Germany, Italy, the UK, and the US.

MwT consists of three parts. The first part focuses on the general issue of AI in society. Cristianini explains how, in an attempt to overcome the technical shortcomings of the 'first generation' AI, the field has accumulated a series of "ethical debts". Because of such debts, the benefits of technical speed and computational power have created several well known social costs (i.e., loss of privacy, unwanted bias, opacity, and so on). Rieder, Simon, and Wong unpack the socio-political values underpinning the current discourse on 'trustworthy AI'.

The second part delves into more specific issues. Hildebrandt provides a taxonomy of the different kinds of bias that machine learning may have. Strandburg talks about the problem of explainability in AI. Machine learning devices are opaque and inscrutable, in the sense that it is not clear how they produce their results. This becomes all the more problematic when such devices are employed as decision-making and decision-supporting tools. Cabitza talks about some unintended consequences of decision-supporting AI devices, whereas Amigoni and Schiaffonati examine the often under-discussed issue of prediction for AI.

The third part opens new avenues for future reflections on the AI-humans interactions. Vaughan and Wallach focus on the concept of intelligibility in the face of the plurality of the societal stakeholders involved in the production and use of AI, and taking into consideration important issues related to social justice, such as fairness and 'democratized' technology. In the concluding chapter, Williamson develops the provoking idea of turning the traditional discourse on the 'ethics of AI' into a discourse on 'the AI of ethics'. What he suggests, in other words, is to consider not only the ethical issues raised by AI but also how AI may influence, or even guide, our moral reasoning.

Each chapter of *MwT* provides both an in-depth state-of-the-art discussion on the most recent developments on the literature on the ethical issues raised by AI as well as original reflections on how to solve, or even to re-frame and re-conceptualise, some of them. This is not an introductory text and its ideal target is that of advanced readers in AI, ethics of innovative technologies, and science policy. While every contribution to the volume offers invaluable food for thought, in what follows I will focus on the chapters that, rather than

discussing trustworthy AI ‘in general’, delve into specific issues raised by concrete forms of AI. The reason for such a choice will become more evident in the concluding section.

3. What could possibly go wrong when we do everything right?

The problem of so-called ‘unintended consequences’ is often discussed in terms of the risk of ‘direct harms’, (i.e., the risk that a new technology may put in peril its users’ health or even life). If the risk of direct harm is higher than the potential benefits, or if the harms are so big that they are not worth the risk, then the research must be stopped, or at least steered toward much safer directions. There is, however, a whole class of unintended consequences that traditional risk/benefit frameworks are unable to capture. These are consequences that are not clearly or directly harmful. These kinds of consequences, in general, are associated with the implementation of innovative technologies that may alter in significant yet unpredictable ways social values, human relations, and cultural dynamics. Social responsibility in innovation and development research, therefore, demands that we consider not only clear-cut and easily quantifiable harms, but also its so-called ‘soft impacts’ (van der Bug 2009).

Cabitza (chapter 6: “Cobra AI: exploring some unintended consequences of our most powerful technology”) employs the suggestive expression of ‘cobra effect’ to discuss how, after solving all the technical problems and reducing the risks of direct harms, some unintended consequences may still ‘bite us back’. He discusses these issues with particular reference to the decision-making and decision-supporting machine learning devices designed for healthcare.

Explainability and bias are regarded as some of the fundamental problems with AI, both in general and in the case of the machine-learning decision-supporting devices in healthcare. Not solving such issues may lead to dangerous direct harms (i.e., AI decision-supporting tools that lead us to biased and bad decisions). Yet, as Cabitza points out, the use of explainable and unbiased AI-devices is no guarantee that good decisions will be taken after all. Completely explainable AI may lead to overconfidence in its decisions. It may also lead to a sort of cognitive laziness, if not ‘digital dementia’, that will represent an obstacle to the decision-making process that AI is supposed to support. Moreover, it is still not clear how decisions will be taken when the output of AI will conflict with the expert opinion of clinicians: trustworthy AI conflicting with trustworthy experts may lead to confusion and undecidability. Beside, many experts are trustworthy even though they are as impenetrable as unexplainable AI: their expertise is the result of a deep and often even ‘tacit’ knowledge and, if asked to give reasons for their decisions, they often offer some ‘rational reconstruction’ which may not mirror their actual reasoning. Yet, interestingly enough, many experts are deemed to be trustworthy and authoritative in virtue of their ‘unexplanability’. Therefore, one may wonder whether explainability is really necessary to trustworthy AI.

Ultimately, what underlies the many ‘cobra effects’ discussed by Cabitza is the uncertainty not of AI per se, but of the ‘AI-humans’ system. While there are many studies about the inefficiencies of human reasoning and the limits of AI, it is still not clear whether a decision-making process distributed across AI and humans will be more effective or will involve less risks.

Cabitza's conclusion may appear as *prima facie* counterintuitive: we should *not* aim at developing a completely trustworthy AI. For instance, we could design AI characterized by 'interaction friction' (i.e., machine-learning devices that do not provide only one output, but different outputs that the user will have to choose among). Instead of focussing on the full explainability and transparency of 'dehumanized' AI tools (that is, AI considered in and by itself), Cabitza invites us to regard 'programmed inefficiencies' as a feature, rather than a bug, of 'human-centred AI', and even to consider the pursuit of an 'AI-decentred humanity', in which AI is considered as an adjunct to the human decision-making process, rather than as one of its essential parts.

4. Trusting AI in the real world

Amigoni and Schiaffonati (chapter 7: "The importance of prediction in designing artificial intelligent systems") tackle a crucial yet under-discussed issue: namely, the fundamental role of prediction for the responsible design and implementation of AI.

The topic of prediction is often overshadowed by discussions about explanation, not only when it comes to the problem of 'explainable AI', but also in more general philosophical debates in philosophy of science. This despite the fact that the original 'Deductive-Nomological model' was a model for *both* scientific explanation *and* scientific prediction (Hempel and Oppenheim 1948). As many pointed out, of course, due to all sorts of uncertainties and because many actual contexts are not governed by universal laws but, rather, by statistical regularities, explanations cannot be as neatly deductive as Hempel hoped. For this reason, many philosophers criticized the D-N model and developed some alternative accounts of explanation. These philosophical developments, however, ended up obscuring the issue of prediction, that was once regarded as being 'symmetrical' with respect to explanation. Only in more recent years prediction is being slowly resurrected philosophers concerned not only with knowledge production, but also with the potential impacts and implications of such knowledge (Douglas 2009, Douglas and Magnum 2013). While the philosophical 'divorce' between explanation and prediction is mirrored in the debates about the ethical issues in AI, the chapter by Amigoni and Schiaffonati mirrors the much needed return of the concept in the debate about socially responsible science and technology.

As the authors rightly point out, explanations are retrospective: what we expect from an explainable AI is indeed the possibility of getting *post hoc* reasons for its results. To trust AI, however, we also need to know that it will behave non-harmfully *ante hoc*. Trustworthy AI, in short, must be both 'explainable' and 'predictable'. Prediction, however, brings up new kinds of uncertainties, related to the so-called problem of 'external validity'. To explain this issue in more concrete terms, the authors discuss the development of autonomous robotics, such as self-driving vehicles. One thing is testing the functioning of an autonomous vehicle in a highly controlled environment. Another thing is being sure that such a vehicle will not be harmful when it will be put in actual roads, under many uncertain conditions (i.e., changing light and weather, sudden reduced or non-optimal visibility conditions, drivers and pedestrians not respecting road rules, and so on). Before deploying it, therefore, AI must be 'experimented' upon.

What the authors have in mind is the kind of experiments conducted in the engineering sciences. Unlike what happens in many natural sciences, in the engineering sciences experiments do not aim at establishing the degree of confidence of a set of hypotheses (unless one is willing to stretch the very concept of a hypothesis in exuberant ways). Rather, engineers make experiments by testing the robustness and behavior of a device, be it mechanical or electronic, under many different conditions and by considering different contingent variables. The authors, in short, maintain that predictable AI could be achieved through what they define as *explorative experiments*, to be conducted in a practical 'socio-technical' context rather than in a highly controlled and 'purified' environment (Amigoni, Reggiani and Schiaffonati 2009, Amigoni and Schiaffonati 2018, Schiaffonati 2022).

Amigoni and Schiaffonati are perhaps too quick in taking for granted the generalizability of explorative experiments in AI. They speak about "the possibility of generalizing the experimental results from one case to similar ones by carefully selecting the settings in which explorative experiments are carried out" (p. 115). However, establishing degrees of similarity between different scenarios, and thus the possibility of generalizing across them, may involve risks of error. Moreover, while some mechanical or technological devices experimented upon in the engineering sciences possess rather stable properties and behaviors to be tested under different varying conditions, AI may learn from different varying conditions and, in turn, vary its own behavior. It would be therefore very difficult to give a conclusive interpretation of the results of an explorative experiment, or even to tell when such an experiment has actually ended.

The authors are obviously aware of these issues, which they tackle in more depth in other writings and that open new avenues for future investigations on how to assess the external validity of autonomous robotics. At the same time, their contribution to the volume has (at least) two strengths: the discussion of the role of prediction for the responsible design of AI, and the establishment of an explicit link between ethical issues, on the one hand, and methodological and epistemological issues, on the other.

5. One, none, and one hundred thousand AIs

Taken on their own, each contribution of *MwT* shed some light on some important socio-ethical issues raised by AI. A more 'comparative' reading of the chapters of the volume, however, would reveal that, for example, Cabitza's proposal to combat the 'cobra effect' of AI in healthcare would hardly work for the autonomous vehicles Amigoni and Schiaffonati talk about. An autonomous vehicle should be trusted for its ability of not causing harm in the face of varying conditions outside a highly controlled environment. If it is not capable of doing so, then the autonomous vehicle would just be unreliable, untrustworthy, and plainly dangerous. Beside, a vehicle displaying 'interaction friction' that a human agent must overcome, as in Cabitza's examples, would not be 'autonomous' to begin with. (Someone accepting the idea of extending Cabitza's arguments about interaction friction for the use of AI in means of transportation could actually conclude that we should abandon, or at least heavily modify, the project of completely autonomous vehicles, and to focus more on the production of AI driving 'support' systems; it is not obvious, however, that the development

of such systems, instead of fully autonomous vehicles, is worth the costs and the efforts.) When it comes to the explorative experiments for testing AI discussed by Amigoni and Schiffonati, it becomes difficult to understand how such experiments could be designed and conducted for the decision-making and decision-supporting AI tools for fields such as healthcare. In those cases, in fact, it is hard to regard such devices as being testable by following an engineering protocol.

Even though their chapters appear in the same volume, and in close succession, it seems that Cabitza and Amigoni and Schiaffonati talk about very different and perhaps even incompatible things. This is not because of some sort of Kuhnian-like 'semantic incommensurability', for which the same term or expression (in this case, 'AI') may mean different things to scientists belonging to conflicting paradigms. Rather, when applied for carrying out specific and concrete tasks, AI may actually become, in a sense, many different things. Indeed, AI used as a support in the decision-making process is not the same thing as AI in autonomous robotics. These different 'AIs' give rise to different kinds of issues that require different perspectives and even different ways of reasoning about them.

On the one hand, this is problematic for both philosophers and policy makers. Oftentimes, philosophy is defined as the study of 'essences', aiming at answering, or even posing, questions about general and fundamental concepts (i.e., "What is Reality?", "What is Justice?", and so on). At least since the classic essay by Martin Heidegger (1954), many philosophers of technology have focussed not only on specific forms of technology, but they have attempted to uncover and analyze its very 'essence' that, presumably, all of its concrete manifestations presumably share. And yet, one is tempted to say that the essence of AI is to change depending on the concrete tool or device it is implemented in, and depending on the task or function it will be employed. Perhaps one of the manifest symptoms of the 'liquid' age we live in, as described by the sociologist Zygmunt Bauman (2000), is the emergence of a 'liquid technology' such as AI, that is ubiquitous and multiform. And, as mentioned, the liquidity of AI is not a problem for philosophers only. The very attempt of establishing rules and providing recommendations for a trustworthy AI 'in general', like those recently advanced by the European Commission, seem destined to run into considerable difficulties. The obvious risk is that the discourse about AI-in-general becomes so detached from the specificity of concrete problems that every guideline would be just vacuous and inapplicable.

On the other hand, however, the reflection stimulated by the critical comparison of different chapters of *Machines we Trust* may also represent an opportunity. While explainability is commonly regarded as one of the fundamental problems, if not 'the' fundamental problem, with AI-in-general, the contribution of Cabitza and Amigoni and Schiaffonati expand the horizons of the debate. Maybe in some contexts the issue of prediction is actually more pressing, if not more fundamental, than that of explainability. Maybe in some other contexts explainability may even lead to undesirable consequences.

Instead of conducting it from general views and principles to concrete applications, the discourse on the socio-ethical issues of AI could proceed from the analysis of specific cases to the establishment of general, although tentative, norms. This way of proceeding may be more difficult and may require a heavier load of intellectual labor to be shared among different kinds

of experts. Ultimately, however, it may produce more insightful results and a finer understanding of where AI, or rather 'AIs', is heading towards.

Funding: This article was written thanks to the generous support of the Beatriu de Pinós Fellowship Scheme (grant reference: 2020 BP 00196).

Conflicts of Interest: The author declares no conflict of interest.

References

Amigoni, F., Reggiani, M. and Schiaffonati, V. (2009) An insightful comparison between experiments in mobile robotics and in science. *Autonomous Robots* 27:313.

Amigoni, F. and Schiaffonati, V. (2018). Ethics for Robots as Experimental Technologies: Pairing Anticipation with Exploration to Evaluate the Social Impact of Robotics, *IEEE Robotics & Automation Magazine* 25:30–36.

Baier, A. (1986) Trust and antitrust. *Ethics* 96:231–260.

Bauman, Z. (2000) *Liquid Modernity.* Cambridge: Polity Press.

Douglas, H. (2009) Reintroducing prediction to explanation. *Philosophy of Science* 76:444–463.

Douglas, H. and Magnus, P. (2013) State of the field: why novel prediction matters. *Studies in History and Philosophy of Science* 44:580–589.

European Commission (2019) *Ethics Guidelines for Trustworthy AI*. https://ec.europa.eu/futurium/en/ai-alliance-consultation Accessed on 26 October 2022.

Heidegger, M. (1954 [1977]). The Question concerning Technology (original title: 'Die Frage nach der Technik'). In Heidegger, *The Question concerning Technology and other essays* (William Lovitt, Trans.). New York: Harper & Row, 1977, pp. 3–35.

Hempel, C. and Oppenheim, P. (1948 [1965]) Studies in the Logic of Explanation, *Philosophy of Science* 15: 135–175.

Holton, R. (1994) Deciding to trust, coming to believe. *Australasian Journal of Philosophy* 72:63–76.

McLeod, C. (2015) Trust. In Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Available online at: https://plato.stanford.edu/archives/fall2015/entries/trust/ Accessed on 26 October 2022.

Schiaffonati, V. (2022) Explorative Experiments: a paradigm shift to deal with severe uncertainty in autonomous robotics. *Perspectives on Science* 30:284–304.

van der Bug, S. (2009) Taking the 'soft impacts' of technology into account: broadening the discourse in research practice. *Social Epistemology* 23:301—16.

Wright, S. (2010) Trust and trustworthiness. *Philosophia* 38:615–627.